

基礎統計学

第5回 分割表の分析

2016. 11. 11

2×2分割表（四分表）

二つの項目 X, Y があり、それぞれ2つの分類 X_1, X_2, Y_1, Y_2 をもつとする。

これらのデータは項目、分類ごとに集計して2×2分割表にまとめられる。

例

X (薬品) : 新薬, 偽薬

Y (効果) : あり, なし

2×2分割表

		Y		
		Y_1	Y_2	計
X	X_1	n_{11}	n_{12}	$N_{1.}$
	X_2	n_{21}	n_{22}	$N_{2.}$
計		$N_{.1}$	$N_{.2}$	N

問題

X, Y には関連があるのかないのか？

独立性の検定

帰無仮説 (H_0) : X, Y には関連がない (独立である)

対立仮説 (H_1) : X, Y には関連がある (独立でない)

として検定を行う.

検定方法

- 表の各マスの数値が5以上のとき χ^2 検定
- 5未満のマスがあるとき Fisher の正確確率検定

ピアソンの χ^2 検定

各マスの期待度数（期待値）を E_{ij} とする.

Table: 期待度数

		Y		
		Y_1	Y_2	計
X	X_1	E_{11}	E_{12}	$N_{1\cdot}$
	X_2	E_{21}	E_{22}	$N_{2\cdot}$
計		$N_{\cdot 1}$	$N_{\cdot 2}$	N

χ^2 値の定義

X, Y には関連がない（独立である）とすると、

$$E_{ij} = \frac{N_{i.} \times N_{.j}}{N}, \quad i, j = 1, 2 \text{ となる.}$$

観測度数 n_{ij} と期待度数 E_{ij} との違いを次式で評価する.

χ^2 値

$$\chi^2 = \sum_{i,j} \frac{(n_{ij} - E_{ij})^2}{E_{ij}}$$

χ^2 分布

Fact

$n_{ij} \geq 5$ のとき, χ^2 は近似的に自由度 1 の χ^2 分布に従う.

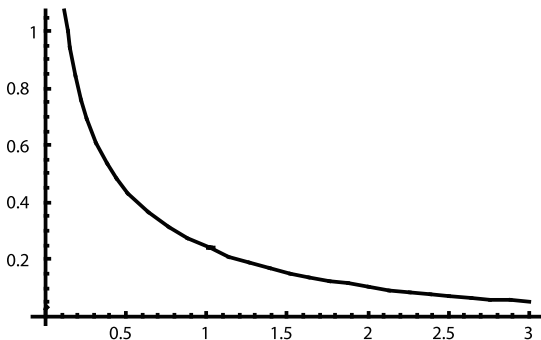


Figure: 自由度 1 の χ^2 分布

判定

- $\chi^2 > \chi_1^2(\alpha)$ のとき (H_0) は棄却
したがって X, Y は独立でない (関連がある).
- $\chi^2 \leq \chi_1^2(\alpha)$ のとき (H_0) は棄却できない.
独立でないとは言えない.

$$\chi_1^2(0.05) = 3.84, \chi_1^2(0.01) = 6.64$$

イエーツ補正

分割表の χ^2 検定は近似的なもので本来の値から多少のずれがある。

そこで0.5の補正をして

イエーツ補正した χ^2 値

$$\chi^2 = \sum_{i,j} \frac{(|n_{ij} - E_{ij}| - 0.5)^2}{E_{ij}}$$

を使うとより近似がよくなることが知られている。
この場合、補正前より p 値は大きくなる。

補足

$k \times l$ の表でも独立性の検定ができる.

この場合は χ^2 値は自由度 $(k-1)(l-1)$ の χ^2 分布に従う.

EZR (R Commander) にて

- 1 medicine.csv を読み込む.
- 2 統計解析 → 名義変数の解析 →

分割表の作成と群間の比率の比較

カイ2乗検定をチェックしておくで、 χ^2 検定が行われる。

Rにて

Rからは

- `data <- read.csv("medicine.csv")`

- クロス集計表をつくる.

```
table <- table(data$Med, data$Eff)
```

- `chisq.test(table)` でイエーツ補正ありの χ^2 検定が行われる.

- `chisq.test(table, correct=FALSE)` でイエーツ補正なしの χ^2 検定が行われる.

Fisher の正確確率検定 (Exact Test)

あるマス数が5未満のときは、 χ^2 値の分布は χ^2 分布からのずれが大きくなり、正確な p 値にならない。

そこで、直接的に確率 (p 値) を計算する方法が正確確率検定である。

注) 一つのマスが5未満のときに Fisher's exact test は使われるが、この方法は5以上のときにも使える手法である。

確率の計算法 (概略)

		Y		
		Y ₁	Y ₂	計
X	X ₁	x	y	n ₁
	X ₂	z	w	n ₂
計		m ₁	m ₂	N

計 n_1, n_2, m_1, m_2 が固定されている状況でデータの組合せで $x = a$ となる確率 $P(x = a)$ を計算すると

$$P(x = a) = \frac{n_1!n_2!m_1!m_2!}{N!a!b!c!d!}$$

となる. ここで $b = n_1 - a (= y)$, $c = m_1 - a (= z)$,
 $d = m_2 - b = m_2 - n_1 + a (= w)$.

判定

有意水準を α とし, $P(x < k) < \alpha/2$ となる最大の k と $P(x > l) < \alpha/2$ となる最小の l を求めると, 棄却域が $x < k$ または $x > l$ となる.

補足: 全てのマス目が5以上の場合で Exact Test を使っても構わない.

再び EZR にて

EZR: 統計解析 → 名義変数の解析 →
分割表の作成と群間の比率の比較
から Exact test が行われる.

R: `fisher.test(table)`

リスク比とオッズ比

あるコホート (cohort) 研究 (リスクファクターあるなしの群を追跡調査する前向きの研究) でつぎの表を得たとする.

		アウトカム		
		あり	なし	計
リスク ファクター	あり	a	b	n_1
	なし	c	d	n_2
計		m_1	m_2	N

リスク比

リスクファクターありの場合の

アウトカムありの確率 $\frac{a}{a+b}$

リスクファクターなしの場合の

アウトカムありの確率 $\frac{c}{c+d}$

リスク比 (Risk Ratio, Relative Risk)

$$RR = \frac{a/(a+b)}{c/(c+d)}$$

これは母集団のリスク比の推定値である。

リスク比の解釈

- リスク比 RR はリスクファクターのあるなしで何倍の危険性があるかを示す数値
- $RR = 1$ ならリスクファクターのあるなしはアウトカムのありなしとは無関係 (独立) ということになる.

リスク比の信頼区間

母集団のリスク比は上記の RR で推定されるが、その 95% 信頼区間はつぎで計算できる。

95% 信頼区間

$$\exp\left(\log RR \pm 1.96 \sqrt{\frac{1}{a} + \frac{1}{b} - \frac{1}{a+c} - \frac{1}{b+c}}\right)$$

$\exp x = e^x$, $\log x =$ 自然対数.

- 1 が信頼区間からはずれていれば有意であると判定できる. ($RR = 1$ は棄却される.)

オッズ比 (Odds Ratio)

ある症例対照 (case-control) 研究の表とする.

		アウトカム		
		あり	なし	計
リスク ファクター	あり	a	b	n_1
	なし	c	d	n_2
計		m_1	m_2	N

症例対照 (case-control) 研究 (アウトカムの情報からリスクファクターを探る後ろ向きの研究) では前記のリスク比を計算することは不適切である. (何故か?)

オッズ

あるイベントが起こる確率を p としたとき, $\frac{p}{1-p}$ をオッズという.

オッズと確率の関係

$$\text{odds} = \frac{p}{1-p}, \quad p = \frac{\text{odds}}{1 + \text{odds}}$$

オッズは確率の別の表現と思える.

- $0 \leq p \leq 1$ だが, $0 \leq \text{odds} \leq \infty$ である.
- $\text{odds} = 1$ であれば起こる起こらないは半々ということの意味する.

四分表のオッズ比

アウトカムありの群でのリスクファクターありのオッズは

$$\frac{a/(a+c)}{c/(a+c)} = \frac{a}{c}$$

アウトカムなしの群でのリスクファクターありのオッズは

$$\frac{b/(b+d)}{d/(b+d)} = \frac{b}{d}$$

オッズ比

2つのオッズの比を**オッズ比**という。(真のオッズ比の推定値)

Odds Ratio (OR)

$$OR = \frac{a/c}{b/d} = \frac{ad}{bc}$$

式の対称性からこれはリスクファクターによるアウトカムありなしのオッズ比に一致する。

アウトカムありの起こる確率が十分小さいとき、 $OR \doteq RR$ となるので**リスク比の代用**とできる。

ORの95%信頼区間

ORの95%信頼区間はつぎで計算できる.

95%信頼区間

$$\exp\left(\log OR \pm 1.96 \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}\right)$$

この場合も1が信頼区間からはずれるとき有意となる.

(母集団のオッズ比は1でない.)

EZRにて

統計解析 → 名義変数の解析 →

分割表の作成と群間の比率の比較でオッズ比の
信頼区間も計算される.